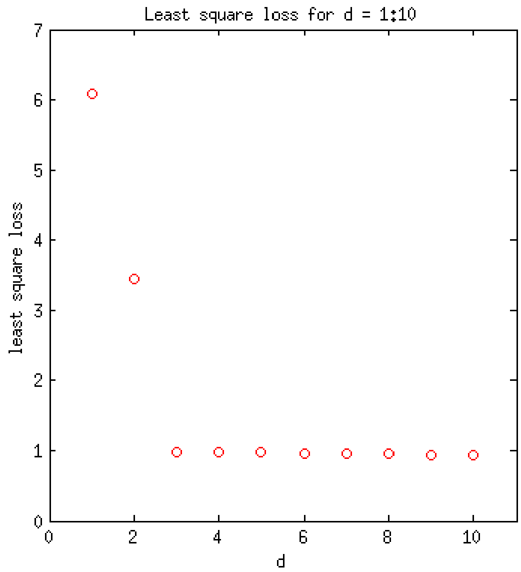
CS 189: Introduction to Machine Learning Spring 2014

Homework 3

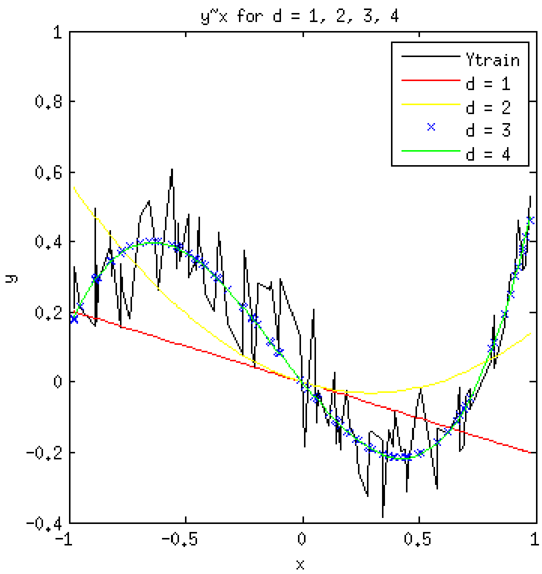
He Ma SID: 22348372

Problem 1

(a) Loss for different ds:



The losses are: 6.0870, 3.4393, 0.9842, 0.9831, 0.9704, 0.9477, 0.9476, 0.9473, 0.9448, 0.9397 for d = 1 to 10.
When the cost function is smallest when d = 10.
d = 10 is not the best choince, because all losses are quite similar for d $\geq$ 3. We can use a smaller d to save computation time and reduce our chance of overfitting.

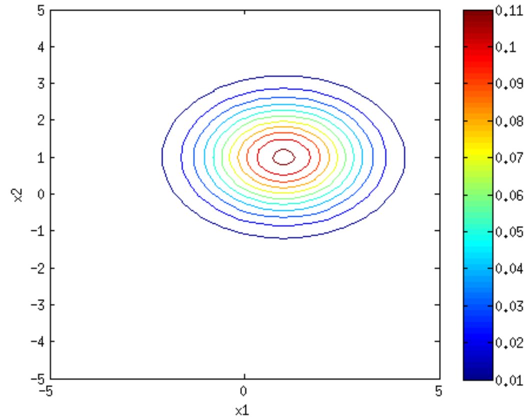


The model for d = 3 and d = 4 looks quite similar for the given set of training points.

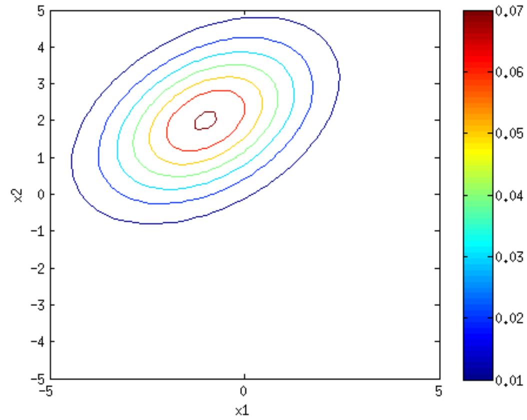
(b) For d = 3, the loss is 5.1013.
For d = 10, the loss is 5.4406.
The model with d = 3 is slightly better than the model with d = 10. It shows that using d = 3 is already pretty good. And the model with d = 10 might overfit the training data.

Problem 2

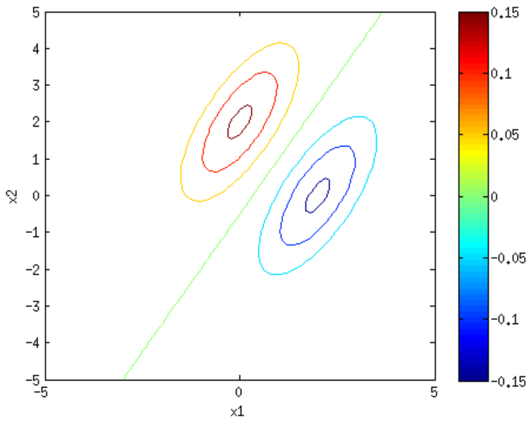
(i) $\mu = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$



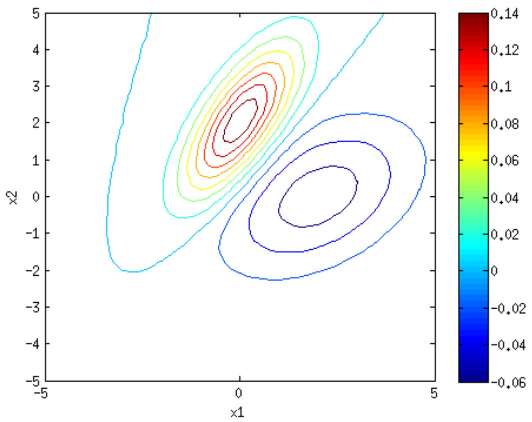
(ii) $\mu = \begin{bmatrix} -1 \\ 2 \end{bmatrix}, \Sigma = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$



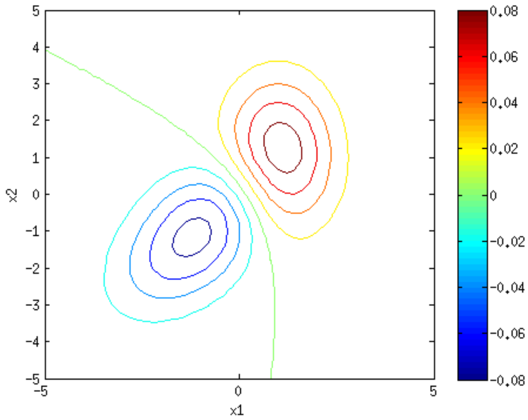
(iii) $\mu1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\mu2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\Sigma1 = \Sigma2 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$



(iv) $\mu 1 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, $\mu 2 = \begin{bmatrix} 2 \\ 0 \end{bmatrix}$, $\Sigma 1 = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$, $\Sigma 2 = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$



(v) $\mu 1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, $\mu 2 = \begin{bmatrix} -1 \\ -1 \end{bmatrix}$, $\Sigma 1 = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$, $\Sigma 2 = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

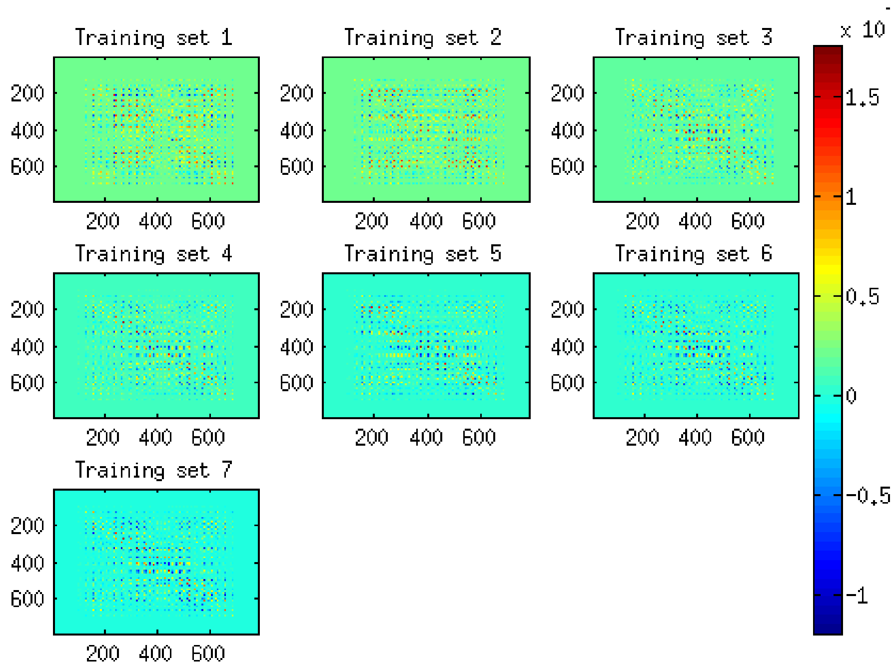


Problem 3

i) From $http://en.wikipedia.org/wiki/Multivariate_normal_distribution$:
$\hat{\mu} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$, which is unbiased.
$\hat{\Sigma} = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$, which is biased.
$\hat{\mu}$ and $\hat{\Sigma}$ are computed in Matlab using the above fomula.
$\hat{\mu}$ is stored in estimated_mus_all in Q3.m at line 11.
estimated_mus_all{i}{j} is the $\hat{\mu}$ for class (j-1) in train set i.
$\hat{\Sigma}$ is stored in estimated_sigmas_all in Q3.m at line 12.
estimated_sigmas_all{i}{j} is the $\hat{\Sigma}$ for class (j-1) in train set i.

ii) Prior probability $P(w = i) = \frac{number\ of\ label\ i\ in\ train\ set}{number\ of\ total\ labels\ in\ train\ set}$
Prior probibilities are stored in prior_dist in Q3.m at line 22.
prior_disti(j) is the prior probability for class (j-1) in train set i

iii) Train sets:

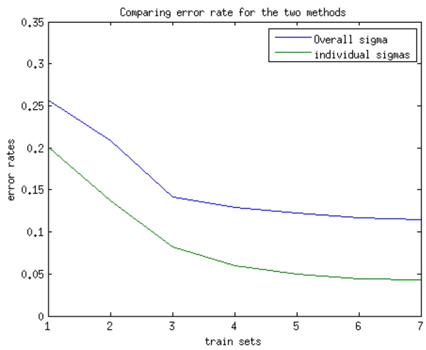


As seen from the plot, for a bigger the training set, the entries which are close to 0 become closer to zero, and the entries which are far from 0 become farther from 0. The pateern of covariance matrix becomes clearer.
We would expect some of these points are more correlated, while others are independent. So it shows that, with a bigger training set, the estimation of covariance matrix becomes more accurate.

iv) (a) Decision boundary:
For a given x, choose class i if:
$exp^{-\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i)}P(w=i) \geq exp^{-\frac{1}{2}(x-\mu_j)^T\Sigma^{-1}(x-\mu_j)}P(w=j)$
for all $i \neq j$.
Error rate for test set using the 7 train sets:
0.2571 0.2081 0.1420 0.1294 0.1226 0.1163 0.1145
Code is in Q3.m. In the code, Σ is added with 0.001 I.

(b) Decision boundary:
For a given x, choose class i if:
$\frac{1}{\sqrt{|\Sigma_i|}}exp^{-\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i)}P(w=i) \geq \frac{1}{\sqrt{|\Sigma_j|}}exp^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)}P(w=j)$
for all $i \neq j$.
Error rate for test set using the 7 train sets:
0.2011 0.1370 0.0825 0.0601 0.0501 0.0438 0.0431
Code is in Q3.m. In the code, each Σ_i is added with 0.001 I. And each Σ is scaled by 960 before computing the determinant.

(c) Comparing the two method:



As seen from the plot, the prediction result using individual Σs is much better than using the oerall Σ.
It implies that each class has quite different Σs so that using $\hat{\Sigma}$ can't really represent all the classes.
That's why the performance is very different.

Problem 4

$J(w,\omega_0) = (y - Xw - \omega_0 1)^T(y - Xw - \omega_0 1) + w^T w$
$J(w,\omega_0) = y^T y + w^T X^T X w + \omega_0^2 n - wy^T Xw - 2y^T\omega_0 1 + 2\omega_0 1^T xw + \lambda w^T w$
$\frac{dJ}{d\omega_0} = 2\omega_0 n - 2y^T 1 = 0$
So $\hat{\omega_0} = \frac{\Sigma y_i}{n} = \bar{y}$

$\frac{dJ}{dw} = 2X^T Xw - 2X^T y + 2X^T 1\omega_0 + 2\lambda w = 0$
$(X^T X + \lambda I)w = X^T y - X^T 1\omega_0$
Because $\bar{x} = 0$, $X^T 1\omega_0 = (\Sigma x_i)\omega_0 = 0$
So $\hat{w} = (X^T X + \lambda I)^{-1} X^T y$

Problem 5

$L(X|\theta) = (\frac{1}{\sqrt{2\pi}\sigma})^n \exp^{-\frac{\Sigma(y_i-\mu_i)^2}{2\sigma^2}}$

$\ln L(X|\theta) = n\ln(\frac{1}{\sqrt{2\pi}\sigma}) - \frac{1}{2\sigma^2}\Sigma(y_i - \omega_0 - \omega_1 x_i)^2$

$\frac{d\ln L(X|\theta)}{d\omega_0} = \frac{1}{2\sigma^2}\Sigma(2(y_i - \omega_0 - \omega_1 x_i)) = 0$

Divide by n: $\bar{y} - \omega_0 - \omega_1\bar{x} = 0$

So $\hat{\omega_0} = \bar{y} - \omega_1\bar{x} \approx E(Y) - \omega_1 E(X)$

$\frac{d\ln L(x|\theta)}{d\omega_1} = \frac{1}{2\sigma^2}\Sigma 2(y_i - \omega_0 - \omega_1 x_i)x_i = 0$

$\Sigma(y_i - \bar{y} + \omega_1\bar{x} - \omega_1 x_i)x_i = 0$

$\Sigma x_i(x_i - \bar{x})\omega_1 = \Sigma(y_i - \bar{y})x_i$ So $\hat{\omega_1} = \frac{\Sigma x_i y_i - \Sigma x_i\bar{y}}{\Sigma x_i^2 - \Sigma x_i\bar{x}_i} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}\bar{x}_i}$

Since $\widehat{cov(X,Y)} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{n} = \frac{\Sigma(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})}{n} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{n}$,

$\widehat{var(X)} = \frac{\Sigma(x_i-\bar{x})^2}{n} = \frac{\Sigma(x_i^2 - 2x_i\bar{x} + \bar{x}^2)}{n} = \frac{\Sigma x_i^2 - n\bar{x}\bar{x}_i}{n}$

So $\hat{\omega_1} = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\Sigma(x_i-\bar{x})^2} = \frac{\Sigma x_i y_i - n\bar{x}\bar{y}}{\Sigma x_i^2 - n\bar{x}\bar{x}_i} \approx \frac{cov(X,Y)}{var(X)}$